Nucleic Acids Research, 2004, Vol. 32, Database issue **D59–D63**
DOI: 10.1093/nar/gkh051

Xpro: database of eukaryotic protein-encoding genes

Vivek Gopalan[1], Tin Wee Tan[1], Bernett T. K. Lee[1] and Shoba Ranganathan[1,2,]*

[1]Department of Biochemistry and [2]Department of Biological Sciences, National University of Singapore, Singapore 119260

Received August 15, 2003; Revised and Accepted September 23, 2003

ABSTRACT

Xpro is a relational database that contains all the eukaryotic protein-encoding DNA sequences contained in GenBank with associated data required for the analysis of eukaryotic gene architecture. In addition to the information found in the GenBank records, which includes properties such as sequence, position, length and description about introns, exons and protein-coding regions, Xpro provides annotations on the splice sites and intron phases. Furthermore, Xpro validates intron positions using alignment information between the record's sequence and EST sequences found in dbEST. In the process of validation, alternative splicing information is also obtained and can be found in the database. The intron-containing genes in the Xpro are also classified as experimental or predicted based on the intron position validation and specific keywords in the GenBank records that are present in predicted genes. An Entrez-like query system, which is familiar to most biologists, is provided for accessing the information present in the database system. A non-redundant set of Xpro database contents is also obtained by cross-referencing to the Swiss-Prot/TrEMBL and Pfam databases. The database currently contains information for 493 983 genes—351 918 intron-containing genes and 142 065 intron-less genes. Xpro is updated for each new GenBank release and is freely available via the internet at http://origin.bic.nus.edu.sg/xpro.

INTRODUCTION

Analysis of the intron/exon features in eukaryotic genes forms the basis for investigating the origin and evolution of genes (1–11). Due to the ever-increasing number of genome and EST sequencing projects, GenBank, the primary repository of nucleotide sequences, is growing at an unprecedented rate. This growth of data and the poor annotation of exon/intron details required for molecular evolution studies in the primary nucleotide database have made specialized databases that provide insight into genomic features specific to eukaryotic genes a necessity.

To facilitate such studies, we have developed Xpro, a relational database system that contains all the eukaryotic protein-coding genes from the GenBank (12) database (release 136). It provides a set of specific data such as exon sequences, the corresponding protein sequences, intron sequences, intron positions and phases, and splice sites, in a well-organized way for analyzing key features of eukaryotic gene architecture. In addition to these, each gene entry in the Xpro database is also cross-referenced to Swiss-Prot/TrEMBL (13) records and the associated Pfam (14) protein families. All these features along with the advanced query system and visual representation of gene structure make this database unique from the currently available databases on eukaryotic gene architecture (15–20).

VALIDATION OF INTRON POSITIONS

The intron positions defined in the GenBank records have to be validated before they are used for any evolutionary studies. This has to be done because GenBank includes not only intron positions derived from experiments but also those based on gene prediction programs such as GeneWise, GENSCAN, etc. Two methods are used for the validation of intron positions. The first one involves text searches of the GenBank header information for the keywords that are commonly used for gene prediction. Those records that contain these keywords are classified as predicted while the rest of the records are classified as experimental as described by Saxonov *et al.* (16). The second method of intron validation is based on the analysis of pairwise alignment of intron-containing mRNA sequences in the Xpro to experimentally derived EST sequences in the dbEST (21) database using BLAT (22). In this method, the intron positions are considered valid only if there is a high level of similarity in the aligned EST sequences in the exon–exon boundaries of the query sequence. Although this method of validation of intron positions is based on experimental EST data and thus is far more accurate than the first method, it does not cover the full data set present in GenBank due to a lack of EST information. These methods are used in tandem and thus provide a higher level of validation than that provided by current databases.

ALTERNATIVE SPLICE VARIANTS

Alternative splicing is a characteristic feature of eukaryotic organisms. It not only increases the complexity and diversity of gene products in the eukaryotes but also forms the basis for analyzing various evolutionary events such as protein domain duplication and exon shuffling. Hence, the Xpro data, which

*To whom correspondence should be addressed. Tel: +65 6874 3566; Fax: +65 6778 2466; Email: shoba@bic.nus.edu.sg represent the gene features in the eukaryotes, are annotated for various types of alternative splicing. The alternative splicing types are assigned based on the gaps that occur in the alignment between the intron-containing GenBank mRNA sequences (only exon sequences) and the EST sequences in dbEST.

INTRON POSITION MAPPING IN PROTEIN HOMOLOGUES

The protein sequences in the Xpro database are redundant as they are derived from GenBank, which does not provide a non-redundant set of protein data. This redundancy is removed by cross-refering the Xpro protein sequences to the Swiss-Prot/TrEMBL database, which contains only non-redundant protein sequences. This results in each Swiss-Prot/TrEMBL entry having multiple Xpro protein sequences. Single representative Xpro protein sequences from each of these clusters thus constitute a set of non-redundant protein sequences. The protein sequences in Xpro database are further classified based on Pfam, into functionally non-redundant protein families. This allows Xpro to aid in the analysis of the conservation and evolution of introns in protein homologues. A graphical display of intron positions mapped to the multiple sequence alignment of Pfam protein family sequences gives further insight to this analysis. In this way, Xpro provides a clean and validated data set targeted for the analysis of eukaryotic gene architecture.

DATABASE CONSTRUCTION AND IMPLEMENTATION

Data source

The data for the Xpro database are obtained from GenBank's invertebrate, plant, primate, rodent and mammalian divisions, which represent all the eukaryotic gene entries. The sequence, length and position data of the introns/exons were obtained by parsing the header, CDS feature and sequence fields in the GenBank records. The intron/exon features and the protein sequence from the GenBank records are extracted respectively from segment details and the translation qualifier present in the CDS field of the feature table. Intron-containing genes are specifically identified based on the keyword 'join' in the CDS field. Exon and intron sequences are derived from the DNA sequence based on the location specified in the CDS join features. In the case of intron-containing genes, if the 3′ and 5′ ends of the flanking exons for an intron are available in different GenBank records—as in the segmented genes—then the intron fragments are derived from DNA sequences corresponding to the reference GenBank locus entries. Further on, phases of the introns are deduced from the exon lengths and the splice sites are extracted from the surrounding exon sequences. Partial sequences are identified based on the '<' or '>' symbol in the CDS field and categorized as 5′ and 3′ deletions accordingly.

Xpro focuses only on the gene architecture of protein-coding regions in eukaryotes. Hence, exons and introns corresponding to 3′ and 5′ UTRs are not included in the database. The pseudogenes, with no protein translations (16), are also not added to the database.

Swiss-Prot/TrEMBL records contain cross-references to GenBank/EMBL protein records and Pfam records. These cross-references are extracted and stored in the relational database for use in the generation of a non-redundant subset of Xpro as described by Schisler and Palmer (17). Pfam accession number, Swiss-Prot/TrEMBL cross-references, domain ranges and multiple sequence alignment information are extracted from the Pfam flat files (version 10.0, July 2003). Each Xpro entry is linked to a Swiss-Prot/TrEMBL entry via the Swiss-Prot/TrEMBL cross-references and using the Swiss-Prot/TrEMBL cross-references extracted from Pfam entries, each Xpro entry is also linked to Pfam entries. This allows individual Xpro records to be classified into protein families and such families can be analyzed for intron distribution within the family.

Data processing

Intron position validation and the capability of intron/exon structures to form alternative splice variants are obtained by alignment of all the mRNA (exon sequences only) of the intron-containing genes in the Xpro database with the EST sequences in the dbEST database using BLAT alignment. A sequence identity cut-off of 90% and the 'fastsearch' option are chosen as BLAT alignment parameters. The BLAT outputs are then stored as tables in the Xpro database for efficient extraction and querying. The intron positions in the Xpro data sets are considered validated if they are covered by at least one EST sequence. The alternative splicing phenomena are analyzed based on gaps and percentage identity in the alignment of the mRNA sequence with EST sequences in dbEST.

Data representation

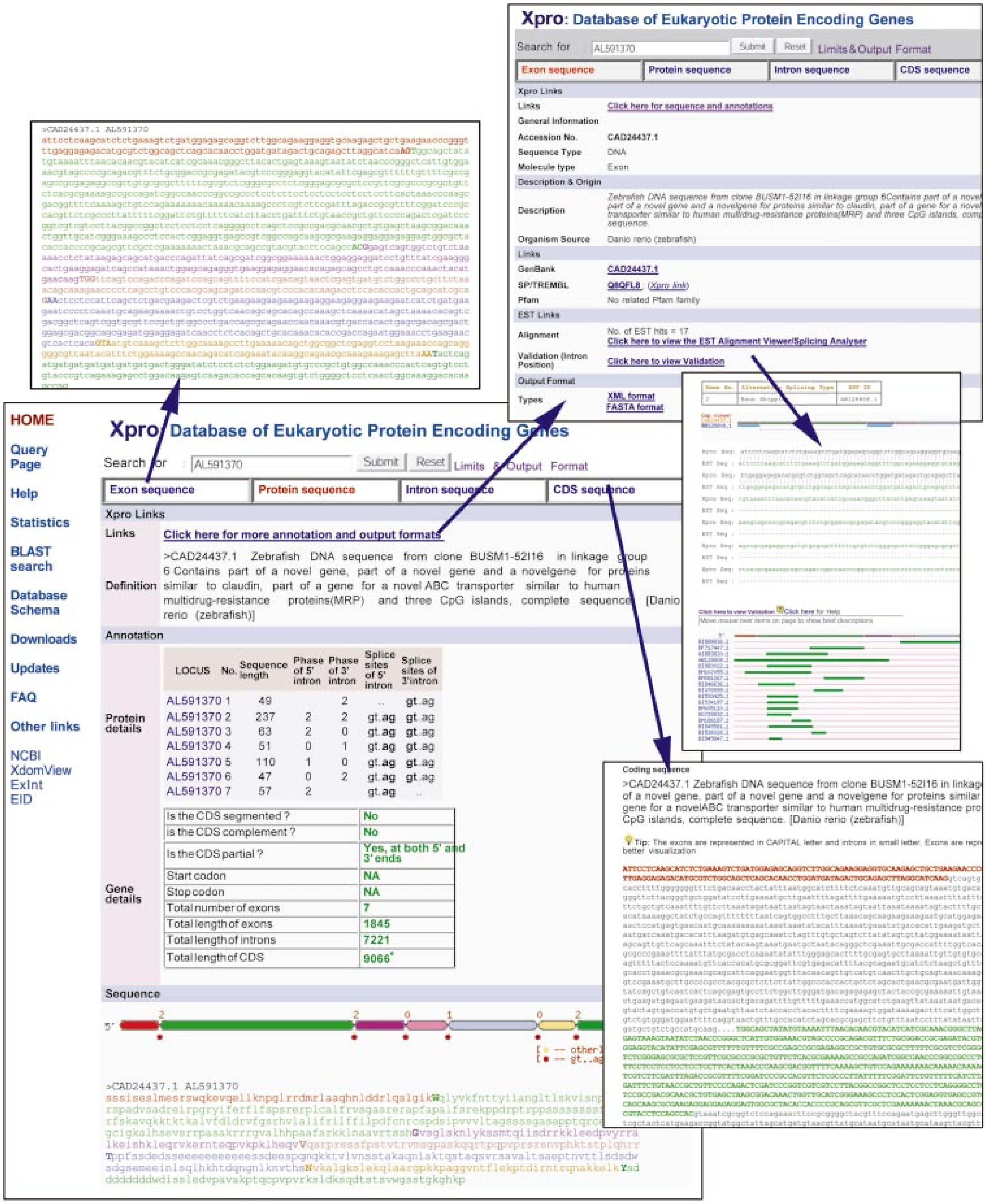
The relational model of data representation is based on the relationships between the locus, GenBank accession, protein accession and intron/exon features present in the GenBank records. The relationship between the GenBank protein accession numbers and accession numbers of the Swiss-Prot/TrEMBL, Pfam and dbEST databases are also considered when building the relational model. The database schema representing the relationship between various tables used in Xpro is available from the Xpro website. The data in Xpro are normalized to remove redundancies and to improve the query speed. The data for the tables in the Xpro database are obtained by parsing the various data sources using PERL scripts.

Implementation

Xpro is housed in a MySQL database (version 3.23.29) (23) running on a UNIX server (SGI ORIGIN 3200, Apache version 1.3). The database is freely available via the internet at http://origin.bic.nus.edu.sg/xpro.

WEB INTERFACE

An Entrez-like query system, familiar to biologists, is provided for efficient search of the data. Thus Xpro can be searched with valid accession numbers, GI numbers, Swiss-Prot/TrEMBL accession numbers, Pfam accession numbers, locus names or keywords from the definition field present in GenBank gene entries.



Xpro: Database of Eukaryotic Protein Encoding Genes

Search for AL591370 Submit Reset Limits & Output Format

Exon sequence | Protein sequence | Intron sequence | CDS sequence

Xpro Links	
Links	Click here for sequence and annotations
General Information	
Accession No.	CAD24437.1
Sequence Type	DNA
Molecule type	Exon
Description & Origin	
Description	*Zebrafish DNA sequence from clone BUSM1-52I16 in linkage group 6Contains part of a novel part of a novel gene and a novelgene for proteins similar to claudin, part of a gene for a novel transporter similar to human multidrug-resistance proteins(MRP) and three CpG islands, comp sequence.*
Organism Source	Danio rerio (zebrafish)
Links	
GenBank	CAD24437.1
SP/TREMBL	Q8QFL8 (Xpro link)
Pfam	No related Pfam family
EST Links	
Alignment	No. of EST hits = 17 Click here to view the EST Alignment Viewer/Splicing Analyser
Validation (Intron Position)	Click here to view Validation
Output Format	
Types	XML format FASTA format

>CAD24437.1 AL591370

HOME

Query Page

Help

Statistics

BLAST search

Database Schema

Downloads

Updates

FAQ

Other links

NCBI

XdomView

ExInt

EID

Xpro: Database of Eukaryotic Protein Encoding Genes

Search for : AL591370 Submit Reset Limits & Output Format

Exon sequence | Protein sequence | Intron sequence | CDS sequence

Xpro Links

Links: Click here for more annotation and output formats

Definition: >CAD24437.1 Zebrafish DNA sequence from clone BUSM1-52I16 in linkage group 6 Contains part of a novel gene, part of a novel gene and a novelgene for proteins similar to claudin, part of a gene for a novel ABC transporter similar to human multidrug-resistance proteins(MRP) and three CpG islands, complete sequence [Danio rerio (zebrafish)]

Annotation

Protein details

LOCUS	No.	Sequence length	Phase of 5' intron	Phase of 3' intron	Splice sites of 5' intron	Splice sites of 3'intron
AL591370	1	49		2	..	gt.ag
AL591370	2	237	2	2	gt.ag	gt.ag
AL591370	3	63	2	0	gt.ag	gt.ag
AL591370	4	51	0	1	gt.ag	gt.ag
AL591370	5	110	1	0	gt.ag	gt.ag
AL591370	6	47	0	2	gt.ag	gt.ag
AL591370	7	57	2		gt.ag	..

Gene details

Is the CDS segmented ?	No
is the CDS complement ?	No
Is the CDS partial ?	Yes, at both 5' and 3' ends
Start codon	NA
Stop codon	NA
Total number of exons	7
Total length of exons	1845
Total length of introns	7221
Total length of CDS	9066*

Sequence

```
>CAD24437.1 AL591370
sssiseslmesrswqkevgellknpglrrdmrlaaqhnlddrlqslgikWqlyvkfnttyiianqltlskvisnp
rspadvsadreirpgryiferflfspsrerplcalfrvsgasrerapfapalfsrekppdrptrppssssssss
rfskavqkktktkalvfdldrvfqsrhvlalifrilffilpdfcnrcspdsipvvvltagssssgasapptqrce
gcigkalhsevsrrpasakrrrgvalhhpaafarkklnaavrtsshGvsglsknlyksmtqiisdrrkkleedpvyrra
lkeishkleqrvkerntegpvkpklhegvVqsrprsssfpstvtrvmsqpaskqqrtpqpvprsrsnvphktstplqhrr
Tppfssdedsseeeeeeeeeeessdeespqmqkktvlvnsstakaqniaktqstaqsvrsavaltsaeptnvttlsdsdw
sdgsemeeinlsqlhkhtdqnqnlknvthsNvkalqkslekqlaargpkkpaggvntflekptdirntrqnakkelkYsd
dddddddwdissledvpavakptqcpvpvrksldksqdtstsvwgsstgkghkp
```

Coding sequence

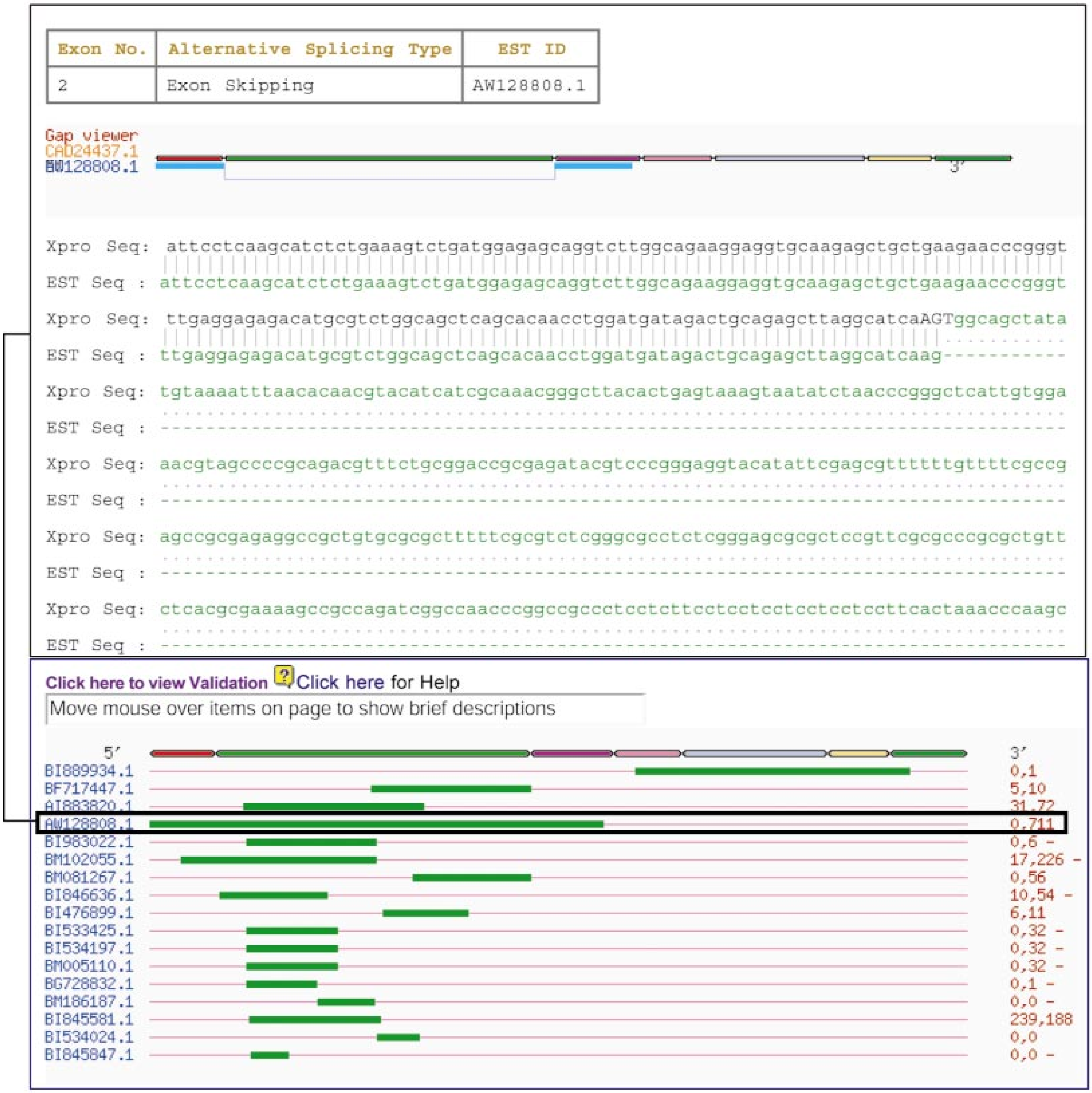
>CAD24437.1 Zebrafish DNA sequence from clone BUSM1-52I16 in linkage of a novel gene, part of a novel gene and a novelgene for proteins similar gene for a novelABC transporter similar to human multidrug-resistance pro CpG islands, complete sequence. [Danio rerio (zebrafish)]

Tip: The exons are represented in CAPITAL letter and introns in small letter. Exons are repr better visualization

Figure 1. Xpro sample results page for the GenBank protein sequence accession number CAAD2447.1.

The query can be limited to a specific data set by enabling the ‘limit & output format’ link in the main menu and by selecting the appropriate conditions listed based on gene type, data set (intron-less or intron-containing or both), cellular location, division and organism. This provision enables the formulation of a more efficient and specific query against Xpro database.

Figure 1 shows the screen shots of the results page and the links in the results page for an Xpro entry. The results are summarized in text, tabular and graphical forms with



Exon No.	Alternative Splicing Type	EST ID
2	Exon Skipping	AW128808.1

```
Xpro Seq: attcctcaagcatctctgaaagtctgatggagagcaggtcttggcagaaggaggtgcaagagctgctgaagaacccgggt
          ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
EST Seq : attcctcaagcatctctgaaagtctgatggagagcaggtcttggcagaaggaggtgcaagagctgctgaagaacccgggt

Xpro Seq: ttgaggagagacatgcgtctggcagctcagcacaacctggatgatagactgcagagcttaggcatcaAGTggcagctata
          |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...........
EST Seq : ttgaggagagacatgcgtctggcagctcagcacaacctggatgatagactgcagagcttaggcatcaag-----------

Xpro Seq: tgtaaaatttaacacaacgtacatcatcgcaaacgggcttacactgagtaaagtaatatctaacccgggctcattgtgga
          ................................................................................
EST Seq : --------------------------------------------------------------------------------

Xpro Seq: aacgtagccccgcagacgtttctgcggaccgcgagatacgtcccgggaggtacatattcgagcgtttttgttttcgccg
          ................................................................................
EST Seq : --------------------------------------------------------------------------------

Xpro Seq: agccgcgagaggccgctgtgcgcgctttttcgcgtctcgggcgcctctcgggagcgcgctccgttcgcgcccgcgctgtt
          ................................................................................
EST Seq : --------------------------------------------------------------------------------

Xpro Seq: ctcacgcgaaaagccgccagatcggccaacccggccgccctcctcttcctcctcctcctcctccttcactaaacccaagc
          ................................................................................
EST Seq : --------------------------------------------------------------------------------
```

Click here to view Validation
Click here for Help
Move mouse over items on page to show brief descriptions
5'
3'

```
BI889934.1   0,1
BF717447.1   5,10
AI883820.1   31,72
AW128808.1   0,711
BI983022.1   0,6 -
BM102055.1   17,226 -
BM081267.1   0,56
BI846636.1   10,54 -
BI476899.1   6,11
BI533425.1   0,32 -
BI534197.1   0,32 -
BM005110.1   0,32 -
BG728832.1   0,1 -
BM186187.1   0,0 -
BI845581.1   239,188
BI534024.1   0,0
BI845847.1   0,0 -
```

Figure 2. Graphical representation of various alternative splicing variants based on the BLAT alignment of the protein-coding region of mRNA sequence derived from the Xpro database with the EST sequences in the dbEST database. Exon sequences in the alignment are represented in different colors so that the type of alternative splicing can be identified.

appropriate links for better understanding and visualization of the gene features.

A unique feature of Xpro is the validation of exon positions based on EST alignments, available from the link marked 'EST View.' An interactive graphical interface for visualizing the BLAT alignment of available ESTs to the query sequence is provided (Fig. 2). By selecting each EST in the alignment, the pairwise alignment is also analyzed for the type of alternative splicing and the sequence alignment itself is displayed in text and graphical formats.

In addition, a BLAST (24) query page is also provided for searching an input sequence against the Xpro database. The output of the BLAST query is displayed graphically with intron positions and alignment gaps mapped on them, allowing better visualization of intron distributions in homologous genes.

AVAILABILITY

The entire Xpro data is freely available for download as FASTA formatted flat files or mysqldump files. The database is freely available via the internet at http://origin.bic.nus.edu.sg/xpro.

Each FASTA formatted sequence in the flat file download is represented by the GenBank definition, protein ID, GI number and accession number in addition to specific gene-structure details such as intron positions, phases, intron–exon size and number and splice sites.

New releases of Xpro will be made available for each new release of GenBank, which is approximately once every 2 months.

XPRO DATA STATISTICS

Detailed statistics of characteristic features of eukaryotic genes in Xpro are provided in the 'statistics' link in the Xpro home page. It includes the distributions of intron lengths, number of introns, phase, splice sites, exon lengths, number of exon and coding sequence lengths for common model organisms such as *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisae.* These distributions are represented as graphs. Other essential features such as GC content, intron density (number of introns/kb coding sequence) and intron penetration (percentage of genes with introns) (17) are also calculated for the model organisms and are presented as tables.

APPLICATIONS

The origin and evolution of introns and their relationship to gene evolution have been analyzed based on eukaryotic gene structure and the properties of genomic elements like introns and exons (1–9). Hence, Xpro can be used as a main data source for gene evolution studies, integrating eukaryotic-specific data present in various diverse data resources.

The validated data sets, the user-friendly web interface and elaborate data statistics make Xpro a unique and valuable database system for analyzing eukaryotic genes.

ACKNOWLEDGEMENTS

The authors thank their colleagues at the Department of Biochemistry, National University of Singapore for their helpful comments, discussions and support. V.G. gratefully acknowledges the National University of Singapore for the award of a scholarship from the Agency for Science, Technology and Research, Singapore (A*STAR).

REFERENCES

1. Gilbert,W. and Glynias,M. (1993) On the ancient nature of introns. *Gene*, **135**, 137–144.
2. Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 901–905.
3. Kriventseva,E.V. and Gelfand,M.S. (1999) Statistical analysis of the exon–intron structure of higher and lower eukaryote genes. *J. Biomol. Struct. Dyn.*, **17**, 281–288.
4. Roy,S.W., Fedorov,A. and Gilbert,W. (2002) The signal of ancient introns is obscured by intron density and homolog number. *Proc. Natl Acad. Sci. USA*, **99**, 15513–15517.
5. Roy,S.W., Fedorov,A. and Gilbert,W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl Acad. Sci. USA*, **100**, 7158–7162.
6. Roy,S.W., Lewis,B.P., Fedorov,A. and Gilbert,W. (2001) Footprints of primordial introns on the eukaryotic genome. *Trends Genet.*, **17**, 496–501.
7. Long,M., Rosenberg,C. and Gilbert,W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA*, **92**, 12495–12499.
8. Vivek,G., Tan,T.W. and Ranganathan,S. (2003) XdomView: protein domain and exon position visualization. *Bioinformatics*, **19**, 159–160.
9. Fedorova,L. and Fedorov,A. (2003) Introns in gene evolution. *Genetica*, **118**, 123–131.
10. Stoltzfus,A., Spencer,D.F., Zuker,M., Logsdon,J.M.,Jr and Doolittle,W.F. (1994) Testing the exon theory of genes: the evidence from protein structure. *Science*, **265**, 202–207.
11. Stoltzfus,A., Spencer,D.F. and Doolittle,W.F. (1995) Methods for evaluating exon–protein correspondences. *Comput. Appl. Biosci.*, **11**, 509–515.
12. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
13. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
14. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
15. Sakharkar,M.K., Kangueane,P., Petrov,D.A., Kolaskar,A.S. and Subbiah,S. (2002) SEGE: A database on 'intron less/single exonic' genes from eukaryotes. *Bioinformatics*, **18**, 1266–1267.
16. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the Exon-Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
17. Schisler,N.J. and Palmer,J.D. (2000) The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res.*, **28**, 181–184.
18. Sakharkar,M., Passetti,F., de Souza,J.E., Long,M. and de Souza,S.J. (2002) ExInt: an Exon Intron database. *Nucleic Acids Res.*, **30**, 191–194.
19. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.*, **24**, 340–341.
20. Lopez,P.J. and Seraphin,B. (2000) YIDB: the Yeast Intron DataBase. *Nucleic Acids Res.*, **28**, 85–86.
21. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
22. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
23. Dubois,P. (2003) *MySQL.* New Riders Press, Indianapolis, IN, USA.
24. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.